Delft University of Technology

Evaluating XAI

A comparison of rule-based and example-based explanations

van der Waa, Jasper; Nieuwburg, Elisabeth; Cremers, Anita; Neerincx, Mark

Important note
To cite this publication, please use the final published version (if applicable).
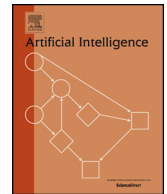Please check the document version above.

Evaluating XAI: A comparison of rule-based and example-based explanations

Jasper van der Waa [a,b,*], Elisabeth Nieuwburg [a,c], Anita Cremers [a], Mark Neerincx [a,b]

[a] *TNO, Perceptual & Cognitive Systems, Soesterberg, Netherlands*
[b] *Technical University of Delft, Interactive Intelligence, Delft, Netherlands*
[c] *University of Amsterdam, Institute of Interdisciplinary Studies, Amsterdam, Netherlands*

A R T I C L E   I N F O

A B S T R A C T

Current developments in Artificial Intelligence (AI) led to a resurgence of Explainable AI (XAI). New methods are being researched to obtain information from AI systems in order to generate explanations for their output. However, there is an overall lack of valid and reliable evaluations of the effects on users' experience of, and behavior in response to explanations. New XAI methods are often based on an intuitive notion what an effective explanation should be. Rule- and example-based contrastive explanations are two exemplary explanation styles. In this study we evaluate the effects of these two explanation styles on system understanding, persuasive power and task performance in the context of decision support in diabetes self-management. Furthermore, we provide three sets of recommendations based on our experience designing this evaluation to help improve future evaluations. Our results show that rule-based explanations have a small positive effect on system understanding, whereas both rule- and example-based explanations seem to persuade users in following the advice even when incorrect. Neither explanation improves task performance compared to no explanation. This can be explained by the fact that both explanation styles only provide details relevant for a single decision, not the underlying rational or causality. These results show the importance of user evaluations in assessing the current assumptions and intuitions on effective explanations.

1. Introduction

Humans expect others to comprehensibly explain decisions that have an impact on them [1]. The same holds for humans interacting with decision support systems (DSS). To help them understand and trust a system's reasoning, such systems need to explain their advice to human users [1,2]. Currently, several approaches are proposed in the field of Explainable Artificial Intelligence (XAI) that allow DSS to generate explanations [3]. Aside from the numerous computational evaluations of implemented methods, literature reviews show that there is an overall lack of high quality user evaluations that add a user-centered focus to the field of XAI [4,5]. As explanations fulfill a user need, explanations generated by a DSS need to be evaluated among these users. This can provide valuable insights into user requirements and effects. In addition, evaluations can be used to benchmark XAI methods to measure the research field's progress.

---

* Corresponding author at: TNO, Perceptual & Cognitive Systems, Soesterberg, Netherlands.
E-mail address: jasper.vanderwaa@tno.nl (J.S. van der Waa).

The contribution of this article is twofold. First, we propose a set of *recommendations* on designing user evaluations in the field of XAI. Second, we performed an extensive user evaluation on the effects of *rule-based* and *example-based* contrastive explanations. The recommendations regard 1) how to construct a theory of the effects that explanations are expected to have, 2) how to select a use case and participants to evaluate that theory, and 3) which types of measurements to use for the theorized effects. These recommendations are intended as a reference for XAI researchers unfamiliar to user evaluations. These recommendations are based on our experience designing a user evaluation and retread knowledge that is more common in fields such as cognitive psychology and Human-Computer Interaction.

The present user study focused on two styles of contrastive explanations and their evaluation. Contrastive explanations in the context of a DSS are those that answer questions as "Why this advice instead of that advice?" [6]. These explanations help users to understand and pinpoint information that caused the system to give one advice over the other. In two separate experiments, we evaluated two contrastive explanation styles. An explanation style defines the way information is structured and is often defined by the algorithmic approach to generate explanations. Note that this is different from explanation form, which defines how it is presented (e.g. textually or visually). The two evaluated styles were *rule-based* and *example-based* explanations, with no explanation as a control. These two styles of explanations are often referred to as means to convey a system's internal workings to a user. However, these statements are not yet formalized into a theory nor compared in detail. Hence, our second contribution is the evaluation of the effects that *rule-based* and *example-based* explanations have on *system understanding* (Experiment I), *persuasive power* and *task performance* (Experiment II). We define system understanding as the user's ability to know how the system behaves in a novel situation and why. The persuasive power of an explanation is defined as its capacity to convince the user to follow the given advice independent of whether it is correct or not. Task performance is defined as the decision accuracy of the combination of the system, explanation and user. Together, these concepts relate to the broader concept of trust, an important topic in XAI research. System understanding is believed to help users achieve an appropriate level of trust in a DSS, and both system understanding and appropriate trust are assumed to improve task performance [7]. Explanations might also persuade the user to various extents, resulting in either appropriate, over- or under-trust, which could affect task performance [8]. Instead of measuring trust directly, we opted for measuring the intermediate variables of understanding and persuasion to better understand how these concepts affect the task.

The way of structuring explanatory information differs between the two explanation styles examined in this study. *Rule-based* explanations are "if... then..." statements, whereas *example-based* explanations provide historical situations similar to the current situation. In our experiments, both explanation styles were *contrastive*, comparing a given advice to an alternative advice that was not given. The *rule-based contrastive explanations* explicitly conveyed the DSS's decision boundary between the given advice and the alternative advice. The *example-based contrastive explanations* provided two examples, one on either side of this decision boundary, both as similar as possible to the current situation. The first example illustrated a situation where the given advice proved to be correct, and the second example showed a different situation where an alternative advice was correct.

Rule-based explanations explicitly state the DSS's decision boundary between the given and the contrasting advice. Given this fact, we hypothesized that these explanations improve a participant's *understanding* of system behavior, causing an improved *task performance* compared to *example-based* explanations. Specifically, we expected participants to be able to identify the most important feature used by the DSS in a given situation, replicate this feature's relevant decision thresholds and use this knowledge to predict the DSS's behavior in novel situations. When the user is confronted with how correct its decisions were, this knowledge would result in a better estimate of when a DSS's advice is correct or not. However, *rule-based* explanations are very factual and provide little information to convince the participant of the correctness of a given advice. As such, we expected *rule-based* explanations to have little *persuasive power*. For the *example-based* explanations we hypothesized opposite effects. As examples of correct past behavior would incite confidence in a given advice, we hypothesized them to hold more *persuasive power*. However, the amount of *understanding* a participant would gain would be limited, as it would rely on participants inferring the separating decision boundary between the examples rather than having it presented to them. Whether persuasive power is desirable in an explanation depends on the use case as well as the performance of the DSS. A low performance DSS combined with a highly persuasive explanation for example, would likely result in a low task performance.

The use case of the user evaluation was based on a diabetes mellitus type 1 (DMT1) self-management context, where patients are assisted by a personalized DSS to decide on the correct dosage of insulin. Insulin is a hormone that DMT1 patients have to administer to prevent the negative effects of the disturbed blood glucose regulation associated with this condition. The dose is highly personal and context dependent, and an incorrect dose can cause the patient short- or long-term harm. The purpose of the DSS's advice is to minimize these adverse effects. This use case was selected for two reasons. Firstly, AI is increasingly more often used in DMT1 self-management [9–11]. Therefore, the results are relevant for research on DSS aided DMT1 self-management. Secondly, this use case was both understandable and motivating for healthy participants without any experience with DMT1. Because DMT1 patients would have potentially confounding experience with insulin administration or certain biases, we recruited healthy participants that imagined themselves in the situation of a DMT1 patient. Empathizing with a patient motivated them to make correct decisions, even if this meant to ignore the DSS's advice in favor of their own choice, or vice versa. This required an understanding of when the DSS's advice would be correct and incorrect and how it would behave in novel situations.

The paper is structured as follows. First we discuss the background and shortcomings of current XAI user evaluations. Furthermore, we provide examples on how *rule-based* and *example-based* explanations are currently used in XAI. The sub-

sequent section describes three sets of recommendations for user evaluations in XAI, based on our experience designing the evaluation as well as on relevant literature. Next, we illustrate our own recommendations by explaining the use case in more detail and offering the theory behind our hypotheses. This is followed by a detailed description of our methods, analysis and results. We conclude with a discussion on the validity and reliability of the results and a brief discussion of future work.

2. Background

The following two sections discuss the current state of user evaluations in XAI and *rule-based* and *example-based* contrastive explanations. The former section illustrates the shortcomings of current user evaluations, formed by either a lack of validity and reliability or the entire omission of an evaluation. The latter discusses the two explanation styles used in our evaluation in more detail, and illustrates their prevalence in the field of XAI.

2.1. User evaluations in XAI

A major goal of Explainable Artificial Intelligence (XAI) is to have AI-systems construct explanations for their own output. Common purposes of these explanations are to increase system understanding [12], improve behavior predictability [13] and calibrate system trust [14,15,8]. Other purposes include support in system debugging [16,12], verification [13] and justification [17]. Currently, the exact purpose of explanation methods is often not defined or formalized, even though these different purposes may result in profoundly different requirements for explanations [18]. This makes it difficult for the field of XAI to progress and to evaluate developed methods.

The difficulties in XAI user evaluations are reflected in recent surveys from Anjomshoae et al. [5], Adadi et al. [19], and Doshi-Velez and Kim [4] that summarize current efforts of user evaluations in the field. The systematic literature review by [5] shows that 97% of the 62 reviewed articles underline that explanations serve a user need but 41% did not evaluate their explanations with such users. In addition, of those papers that performed a user evaluation, relatively few provided a good discussion of the context (27%), results (19%) and limitations (14%) of their experiment. The second survey from [19] evaluated 381 papers and found that only 5% had an explicit focus on the evaluation of the XAI methods. These two surveys show that, although user evaluations are being conducted, many of them provide limited conclusions for other XAI researchers to build on.

A third survey by [4] discusses an explicit issue with user evaluations in XAI. The authors argue to systematically start evaluating different explanations styles and forms in various domains, a rigor that is currently lacking in XAI user evaluations. To do so in a valid way, several recommendations are given. First, the application level of the study context should be made clear; either a real, simplified or generic application. Second, any (expected) task-specific explanation requirements should be mentioned. Examples include the average human level of expertise targeted, and whether the explanation should address the entire system or a single output. Finally, the explanations and their effects should be clearly stated together with a discussion of the study's limitations. Together, these three surveys illustrate the shortcomings of current XAI user evaluations.

From several studies that do focus on evaluating user effects, we note that the majority focuses on subjective measurement. Surveys and interviews are used to measure user satisfaction [20,21], the goodness of an explanation [22], acceptance of the system's advice [23,24] and trust in the system [25–28]. Such subjective measurements can provide a valuable insight in the user's perspective on the explanation. However, these results do not necessarily relate to the behavioral effects an explanation could cause. Therefore, these subjective measurements require further investigation to see if they correlate with a behavioral effect [7]. Without such an investigation, these subjective results only provide information on the user's beliefs and opinions, but not on actual gained understanding, trust or task performance. Some studies, however, do perform objective measurements. The work from [29] for example, measured both subjective ease-of-use of an explanation and a participant's capacity to correctly make inferences based on the explanations. This allowed the authors to differentiate between behavioral and self-perceived effects of an explanation, underlining the value of performing objective measurements.

The above described critical view on XAI user evaluations is related to the concepts of construct validity and reliability. These two concepts provide clear standards to scientifically sound user evaluations [30–32]. The construct validity of an evaluation is its accuracy in measuring the intended constructs (e.g. understanding or trust). Examples of how validity may be harmed is a poor design, ill defined constructs or arbitrarily selected measurements. Reliability, on the other hand, refers to the evaluation's internal consistency and reproducibility, and may be harmed by a lack of documentation, an unsuitable use case or noisy measurements. In the social sciences, a common condition for results to be generalized to other cases and to infer causal relations is that a user evaluation is both valid and reliable [30]. This can be (partially) obtained by developing different types of measurements for common constructs. For example, self-reported subjective measurements such as ratings and surveys can be supplemented by behavioral measurements to gather data on the performance in a specific task.

2.2. Rule-based and example-based explanations

Human explanations tend to be contrastive: they compare a certain phenomenon (fact) with a hypothetical one (foil) [33,34]. In the case of a decision support systems (DSS), a natural question to ask is "Why this advice?". This question
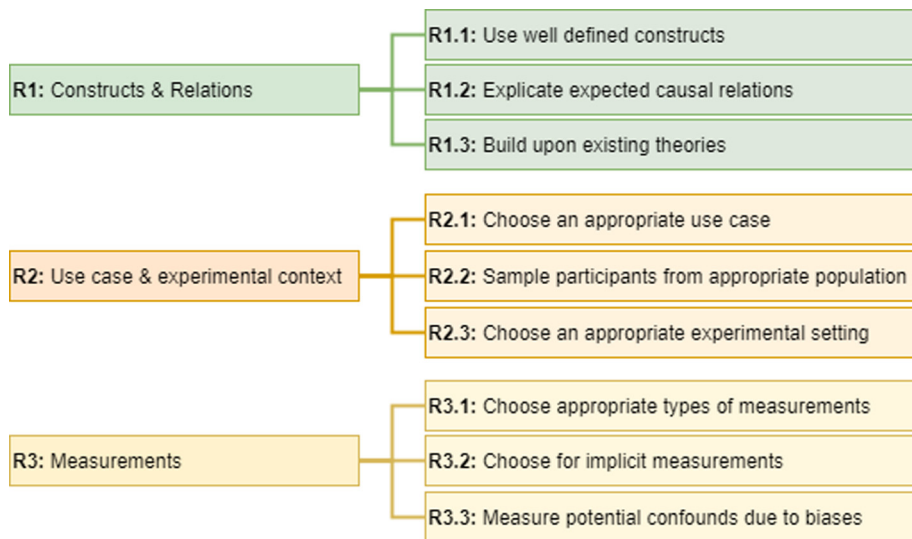
Fig. 1. An overview of three sets of practical recommendations to improve user evaluations for XAI.

implies a contrast, as the person asking this question often has an explicit contrasting foil in mind. In other words, the implicit question is "Why this advice and not that advice?". The specific contrast allows the explanation to be limited to the differences between fact and foil. Humans use contrastive explanations to explain events in a concise and specific manner [2]. This advantage also applies to systems: contrastive explanations narrow down the available information to a concrete difference between two outputs.

Contrastive explanations can vary depending on the way the advice is contrasted with a different advice, for example using rules or examples. Within the context of a DSS advising an insulin dose for DMT1 self-management, a contrastive rule-based explanation could be: "Currently the temperature is below 10 degrees and a lower insulin dose is advised. If the temperature was above 30 degrees, a normal insulin dose would have been advised." This explanation contains two rules that explicitly state the differentiating decision boundaries between the fact and foil. Several XAI methods aim to generate this type of "if... then..." rules, such as [35–38].

An example-based explanation refers to historical situations in which the advice was found to be true or false: "The temperature is currently 8 degrees, and a lower insulin dose is advised. Yesterday was similar: it was 7 degrees and the same advice proved to be correct. Two months ago, when it was 31 degrees, a normal dose was advised instead, which proved to be correct for that situation". Such example- or instance-based explanations are often used between humans, as they illustrate past behavior and allow for generalization to new situations [39–42]. Several XAI methods try to identify examples to generate such explanations, for example those from [43–47].

Research on system explanations using rules and examples is not new. Most of the existing research focused on exploring how users preferred a system would reason, by rules or through examples. For example, users prefer an example-based spam-filter over a rule-based [48], while they prefer spam-filter explanations to be rule-based [49]. Another evaluation showed that the number of rule factors in an explanation had an effect on task performance by either promoting system over-reliance (too many factors) or self-reliance (too few factors) [50]. Work by Lim et al. [51] shows that rule-based explanations cause users to understand system behavior, especially if those rules explain why the system behaves in a certain way as opposed to why it does not behave in a different (expected) way. Studies such as these tend to evaluate either rules or examples, depending on the research field (e.g. recommender system explanations tend to be example-based) but few compare rules with examples.

3. Recommendations for XAI user evaluations

As discussed in Section 2.1, user evaluations play an invaluable role in XAI but are often omitted or of insufficient quality. Our main contribution is a thorough evaluation of rule-based and example-based contrastive explanations. In addition, we believe that the experience and lessons learned in designing this evaluation can be valuable for other researchers. Especially researchers in the field of XAI that are less familiar with user evaluations can benefit from guidance in the design of user studies incorporating knowledge from different disciplines. To that end, we propose three sets of recommendations with practical methods to help improve user evaluations. An overview is provided in Fig. 1.

3.1. R1: Constructs and relations

As stated in Section 2.1, the field of XAI often deals with ambiguously defined concepts such as 'understanding'. We believe that this hinders the creation and replication of XAI user evaluations and their results. Through clear definitions and motivation, the contribution of the evaluation becomes more apparent. This also aids other researchers to extend on the results. We provide three practical recommendations to clarify the evaluated constructs and their relations.

Our first recommendation is to clearly define the intended purposes of an explanation in the form of a construct. A construct is either the intended purpose, an intermediate requirement for the purpose or a potential confound to your purpose. Constructs form the basis of the scientific theory underlying XAI methods and user evaluations. By defining a construct, it becomes easier to develop measurements. Second, we recommend to clearly define the relations expected between the constructs. A concrete and visual way to do so is through a Causal Diagram which presents the expected causal relations between constructs [52]. These relations form your hypotheses and make sure they are formulated in terms of your constructs. Clearly stating hypotheses allows other researchers to critically reflect on the underlying theory assumed, proved or falsified with the evaluation. It offers insight in how constructs are assumed to be related and how the results support or contradict these relations.

Our final recommendation regarding constructs is to adopt existing theories, such as from philosophy, (cognitive) psychology and from human-computer interaction (see [2,6] for an overview). The former provides construct definitions whereas the latter two provide theories of human-human and human-computer explanations. These three recommendations to define constructs and their relations and grounding them in other research disciplines can contribute to more valid and reliable user evaluations. In addition, this practice allows results to be meaningful even if hypotheses are rejected, as they falsify a scientific theory that may have been accepted as true.

3.2. R2: Use case and experimental context

The second set of recommendations regards the experimental context, including the use case. The use case determines the task, the participants that can and should be used, the mode of the interaction, the communication that takes place and the information available to the user [53]. As [4] already stated, the selected use case has a large effect on the conclusions that can be drawn and the extent to which they can be generalized. Also, the use case does not necessarily need to be of high fidelity, as a low fidelity allows for more experimental control and a potentially more valid and reliable evaluation [54]. We recommend to take these aspects into account when determining the use case and to reflect on the choices made when interpreting the results the user evaluation. This improves both the validity and reliability of the evaluation. A concrete way to structure the choice for a use case is to follow the taxonomy provided by [4] (see Section 2.1) or a similar one.

The second recommendation concerns the sample of participants selected, as this choice determines the initial knowledge, experience, beliefs, opinions and biases the users have. Whether participants are university students, domain experts or recruited online through platforms such as Mechanical Turk, the characteristics of the group will have an effect on the results. The choice of population should be governed by purpose of the evaluation. For example, our evaluation was performed with healthy participants rather than diabetes patients, as the latter tend to vary in their diabetes knowledge and suffer from misconceptions [55]. These factors can interfere in an exploratory study such as ours, in which the findings are not domain specific. Hence, we recommend to invest in both understanding the use case domain and reflecting on the intended purpose of the evaluation. These considerations should be consolidated in inclusion criteria to ensure that the results are meaningful with respect to the study's aim.

Our final recommendation related to the context considers the experimental setting and surroundings, as these may affect the quality and generalizability of the results. An online setting may provide a large quantity of readily available participants, but the results are often of ambiguous quality (see [56] for a review). If circumstances allow, we recommend to use a controlled setting (e.g. a room with no distractions, or a use case specific environment). This allows for valuable interaction with participants while reducing potential confounds that threaten the evaluation's reliability and validity.

3.3. R3: Measurements

Numerous measurements exist for computational experiments on suggested XAI methods (for example; fidelity [57], sensitivity [58] and consistency [59]). However, there is a lack of validated measurements for user evaluations [7]. Hence, our third group of recommendations regards the type of measurement to use for the operationalization of the constructs. We identify two main measurement types useful for XAI user evaluations: self-reported measures and behavioral measures. Self-reported measures are subjective and are often used in XAI user evaluations. They provide insights in users' conscious thoughts, opinions and perceptions. We recommend the use of self-reported measures for subjective constructs (e.g. perceived understanding), but also recommend a critical perspective on whether the measures indeed address the intended constructs. Behavioral measures have a more observational nature and are used to measure actual behavioral effects. We recommend their usage for objectively measuring constructs such as understanding and task performance. Importantly however, such measures often only measure one aspect of behavior. Ideally, a combination of both measurement types should be used to assess effects on both the user's perception and behavior. In this way, a complete perspective on a construct can be

obtained. In practice, some constructs lend themselves more for self-reported measurements, for example a user's perception on trust or understanding. Other constructs are more suitable for behavioral measurements, such as task performance, simulatability, predictability, and persuasive power.

Furthermore, we recommend to measure explanation effects implicitly, rather than explicitly. When participants are not aware of the evaluation's purpose, their responses may be more genuine. Also, when measuring understanding or similar constructs, the participant's explicit focus on the explanations may cause skewed results not present in a real world application. This leads to our third recommendation to measure potential biases. Biases can regard the participant's overall perspective on AI, the use case, decision-making or similar. However, biases can also be introduced by the researchers themselves. For example, one XAI method can be presented more attractively or reliably than another. It can be difficult to prevent such biases. One way to mitigate these biases is to design how the explanation are presented, the explanation form, in an iterative manner with expert reviews and pilots. In addition, one can measure these biases nonetheless if possible and reasonable. For example, a usability questionnaire can be used to measure potential differences between the way explanations are presented in the different conditions. For our study we designed the explanations iteratively and verified that the chosen form for each explanation type did not differ significantly in the perception of the participants.

4. The use case: diabetes self-management

In this study, we focused on personalized healthcare, an area in which machine learning is promising and explanations are essential for realistic applications [60]. Our use case is that of assisting patients with diabetes mellitus type 1 (DMT1) with personalized insulin advice. DMT1 is a chronic autoimmune disorder in which glucose homeostasis is disturbed and intake of the hormone insulin is required to balance glucose levels. Since blood glucose levels are influenced by both environmental and personal factors, it is often difficult to find the adequate dose of insulin that stabilizes blood glucose levels [61]. Therefore, personalized advice systems can be a promising tool in DMT1 management to improve quality of life and mitigate long-term health risks.

In our context, a DMT1 patient finds it difficult to find the optimal insulin dose for a meal in a given situation. On the patient's request, a fictitious intelligent DSS provides assistance with the insulin intake before a meal. Based on different internal and external factors (e.g. hours of sleep, temperature, past activity, etc.), the system may advise to take a normal insulin dose, or a higher or lower dose than usual. For example, the system could advise a lower insulin dose based on the current temperature. The factors that were used in the evaluation are realistic, and were based on Bosch [62] and an interview with a DMT1 patient.

In this use case, both the advice and the explanations are simplified. This study therefore falls under the human grounded evaluation category of Doshi-Velez and Kim [4]: a simplified task of a real-world application. The advice is binary (higher or lower), whereas in reality one would expect either a specific dose or a range of suggested doses. This simplification allowed us to evaluate with novice users (see Section 6.3), as we could limit our explanation to the effects of a too low or too high dosage without going into detail about effects of specific doses. Furthermore, this prevented the unnecessary complication of having multiple potential foils for our contrastive explanations. Although the selection of the foil, either by system or user, is an interesting topic regarding contrastive explanations, it was deemed out of scope for this evaluation. The second simplification was that the explanations were not generated using a specific XAI method, but designed by the researchers instead. Several design iterations were conducted based on feedback from XAI researchers and interaction designers to remove potential design choices in the explanation form that could cause one explanation to be favored over another. Since the explanations were not generated by a specific XAI method, we were able to explore the effects of more prototypical rule- and example-based explanations inspired by multiple XAI methods that generate similar explanations (see Section 2.2).

There are several limitations caused by these two simplifications. First, we imply that the system can automatically select the appropriate foil for contrastive explanations. Second, we assume that the XAI method is able to identify only the most relevant factors to explain a decision. Although this assumes a potentially complex requirement for the XAI method, it is a reasonable assumption as humans prefer a selective explanation over a complete one [2].

5. Constructs, expected relations and measurements

The user evaluation focused on three constructs: system understanding, persuasive power, and task performance. Although an important goal of offering explanations is to allow users to arrive at the appropriate level of trust in the system [63,7], the construct of trust is difficult to define and measure [18]. As such, our focus was on constructs influencing trust that were more suitable to translate into measurable constructs; the intermediate construct of system understanding and the final construct of task performance of the entire user-system combination. The persuasive power of an explanation was also measured, as an explanation might cause over-trust in a user; believing that the system is correct while it is not, without having a proper system understanding. As such, the persuasive power of an explanation confounds to the effect of understanding on task performance.

Both *contrastive rule-* and *example-based* explanations were compared to each other with *no explanation* as a control. Our hypotheses are visualized in a Causal Diagram depicted in Fig. 2 [52]. From *rule-based* explanations we expected participants to gain a better understanding of when and how the system arrives at a specific advice. *Contrastive rule-based* explanations
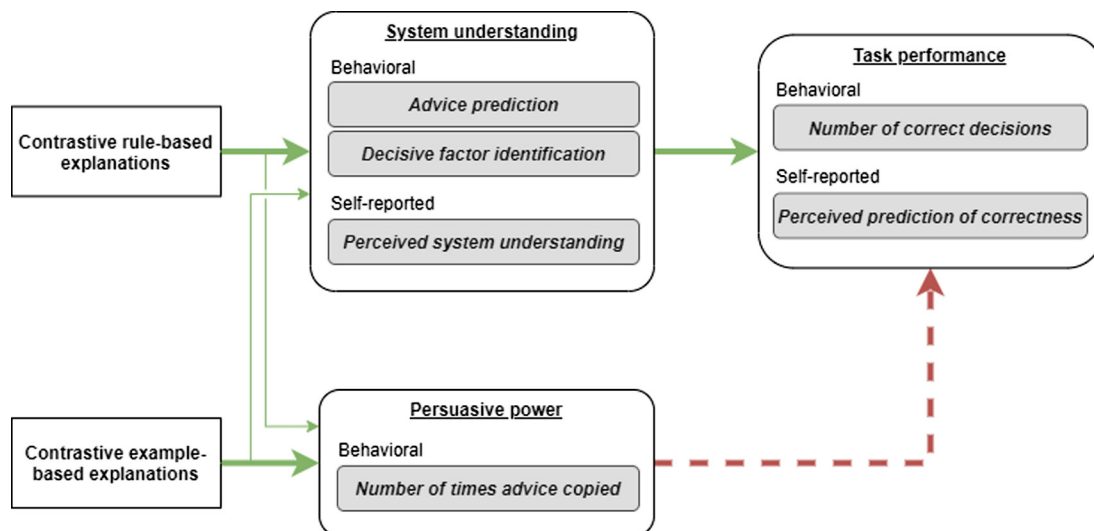
Fig. 2. Our theory, depicted as a Causal Diagram. It describes the expected effects of contrastive rule- and example-based explanations on the constructs of system understanding, persuasive power and task performance. The solid and green arrows depict expected positive effects and the dashed and red arrow depicts a negative effect. The arrow thickness depicts the size of the expected effect. The opaque grey boxes are the measurements that were performed for that construct, divided into behavioral and self-reported measurements.

explicate the system's decision boundary between fact and foil and we expected the participants to recall and apply this information. Second, we expected that *contrastive example-based* explanations persuade participants to follow the advice more often. We believe that examples raise confidence in the correctness of an advice as they illustrate past good performance of the system. Third, we hypothesized that both system understanding and persuasive power have an effect on task performance. Whereas this effect was expected to be positive for system understanding, persuasive power was expected to affect task performance negatively in case a system's advice is not always correct. This follows the argumentation that persuasive explanations can cause harm as they may convince users to over-trust a system [64]. Note that we conducted two separate experiments to measure the effects of an explanation type on understanding and persuasion. This allowed us to measure the effect of each construct separately on task performance, but not their combined effect (e.g. whether sufficient understanding can counteract the persuasiveness of an explanation).

The construct of understanding was measured with two behavioral measurements and one self-reported measurement. The first behavioral measurement assessed the participant's capacity to correctly *identify the decisive factor* of the situations in the system's advice. This measured to what extent the participant recalled what factor the system believed to be important for a specific advice and situation. Second, we measured the participant's ability to accurately *predict the advice* in novel situations. This tested whether the participant obtained a mental model of the system that was sufficiently accurate enough to predict its behavior in novel situations. The self-reported measurement tested the participant's *perceived system understanding*. This provided insight in whether participants over- or underestimated their understanding of the system compared to what their behavior told us.

Persuasive power of the system's advice was measured with one behavioral measurement, namely the number of times participants *copied the advice*, independent of its correctness. If participants that received an explanation followed the advice more often than participants without an explanation, we addressed this to the persuasiveness of the explanation.

Task performance was measured as the *number of correct decisions*, a behavioral measurement, and *perception of predicting advice correctness*, a self-reported measurement. We assumed a system that did not have a 100% accurate performance, meaning that it also made incorrect decisions. Therefore, the number of correct decisions made by the participant while aided by the system could be used to measure task performance. The self-reported measure allowed us to measure how well participants believed they could predict the correctness of the system advice.

Finally, two self-reported measurements were added to check for potential confounds. The first was a brief *usability questionnaire* addressing issues such as readability and the organization of information. This could reveal whether one explanation style was designed and visualized better than the other, which would be a confounding variable. The second, *perceived system accuracy*, measured how accurate the participant thought the system was. This could help identify a potential over- or underestimation of the usefulness of the system, that could have affected to what extent participants attended to the system's advice and explanation.

The combination of self-reported and behavioral measurements enabled us to draw relations between our observations and a participant's own perception. Finally, by measuring a single construct with different measurements (known as triangulation [65]) we could identify and potentially overcome biases and other weaknesses in our measurements.

Current situation

Planned alcohol intake
3 units

Water intake so far
5 glasses

Hours slept
6 hours

The system advises a
lower dose of insulin

Your planned alcohol intake is more than 1 unit.

If this would have been 1 unit or less, the system would have advised a normal dose.

Comparable situation from your past

Planned alcohol intake
3 units

Water intake so far
5 glasses

Hours slept
7 hours

The system advises a
lower dose of insulin

Here, your planned alcohol intake was 3 units and the system also advised a lower dose of insulin.

That advice had a positive effect on your blood sugar level.

Current situation

Planned alcohol intake
3 units

Water intake so far
5 glasses

Hours slept
6 hours

The system advises a
lower dose of insulin

Comparable situation from your past

Planned alcohol intake
1 unit

Water intake so far
4 glasses

Hours slept
6.5 hours

The system advises a
normal dose of insulin

Here, your planned alcohol intake was 1 unit and the system advised a normal dose of insulin instead.

That advice had a positive effect on your blood sugar level.

Fig. 3. The contrastive rule-based (above) and example-based (below) explanation styles. Participants could view the situation, advice and explanation indefinitely.

6. Methods

In this section we describe the operationalization of our user evaluation in two separate experiments in the context of DSS advice in DMT1 self-management (see Section 4). *Experiment I* focused on the construct of system understanding. *Experiment II* focused on the constructs of persuasive power and task performance. The explanation style (contrastive rule-based, contrastive example-based or no explanation) was the independent variable in both experiments and was tested between-subjects. See Fig. 3 for an example of each explanation style.

The experimental procedure was similar in both experiments:

1. Introduction. Participants were informed about the study, use-case and task, as well as presented with a brief narrative about a DMT1 patient for immersive purposes.
2. Demographics questionnaire. Age and education level were inquired to identify whether the population sample was sufficiently broad.
3. Pre-questionnaire. Participants were questioned on DMT1 knowledge to assess if DMT1 was sufficiently introduced and to check our assumption that participants had no additional domain knowledge.

Fig. 4. A schematic overview of the learning (left) and testing (right) block in Experiment I.
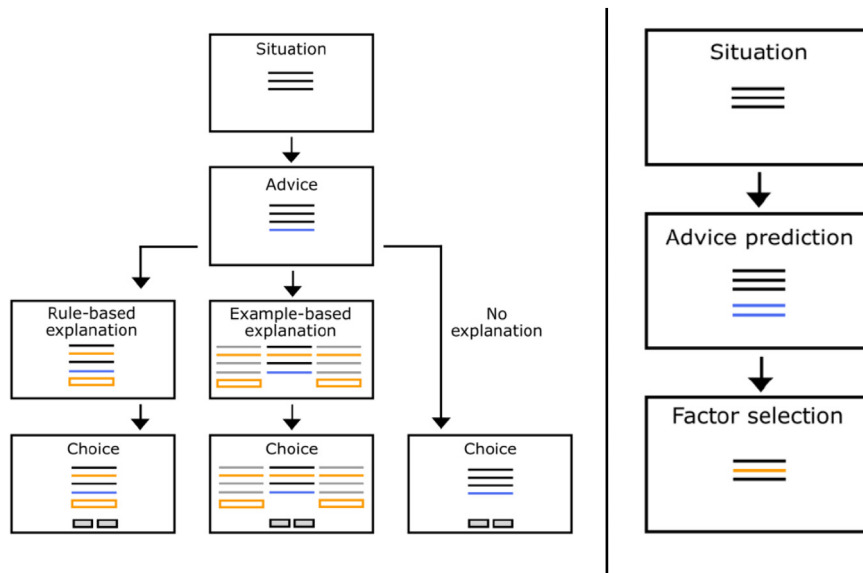
Table 1
An overview of the nine factors that played a role in the experiment. For each factor, its influence on the correct insulin dose is shown, as well as the system threshold for that influence. The thresholds differed between the two experiments and the set of rules of the first experiment were defined as the ground truth. Three factors served as fillers and had no influence.

Factor	Insulin dose	Exp. I Rules	Exp. II Rules
Planned alcohol intake	Lower dose	>1 unit	>1 unit
Planned physical exercise	Lower dose	>17 minutes	>20 minutes
Physical health	Lower dose	Diarrhoea & Nausea	Diarrhoea & Nausea
Hours slept	Higher dose	<6 hours	<6 hours
Environmental temperature	Higher dose	>26 °C	>31 °C
Anticipated tension level	Higher dose	>3 (a little tense)	>4 (quite tense)
Water intake so far	-	-	-
Planned caffeine intake	-	-	-
Mood	-	-	-

4. Learning block. Multiple stimuli were presented, accompanied with either the example- or rule-based explanations, or no explanations (control group).
5. Testing block. Several trials followed to conduct the behavioral measurements (*advice prediction* and *decisive factor identification* in *Experiment I*, the *number of times advice copied* and *number of correct decisions* in *Experiment II*).
6. Post-questionnaire. A questionnaire was completed to obtain self-reported measurements (*perceived system understanding* in *Experiment I* and *perceived prediction of advice correctness* in *Experiment II*).
7. Usability questionnaire. Participants filled out a usability questionnaire to identify potential interface related confounds.
8. Control questionnaire. The experimental procedure concluded with several questions to assess whether the purpose of the study was suspected and to measure *perceived system accuracy* to identify over- or under-trust in the system.

6.1. Experiment I: System understanding

The purpose of *Experiment I* was to measure the effects of *rule-based* and *example-based* explanations on system understanding compared to each other and to the control group with *no explanations*. See Fig. 4 for an overview of both the learning and testing blocks. The learning block consisted of 18 randomly ordered trials, each trial describing a single situation with three factors and values from Table 1. The situation description was followed by the system's advice, in turn followed by an explanation (in the experimental groups). Finally, the participant was asked to make a decision on administering a higher or lower insulin dose than usual. This block served only to familiarize the participant with the system's advice and its explanation and to learn when and why a certain advice was given. Participants were not instructed to focus on the explanations in the learning block, nor were they informed of the purpose of the two blocks.

In the testing block, two behavioral measures were used to test the construct of understanding: *advice prediction* and *decisive factor identification*. The testing block consisted of 30 randomized trials, each with a novel situation description. Each description was followed by the question what advice the participant thought the system would give. This formed the
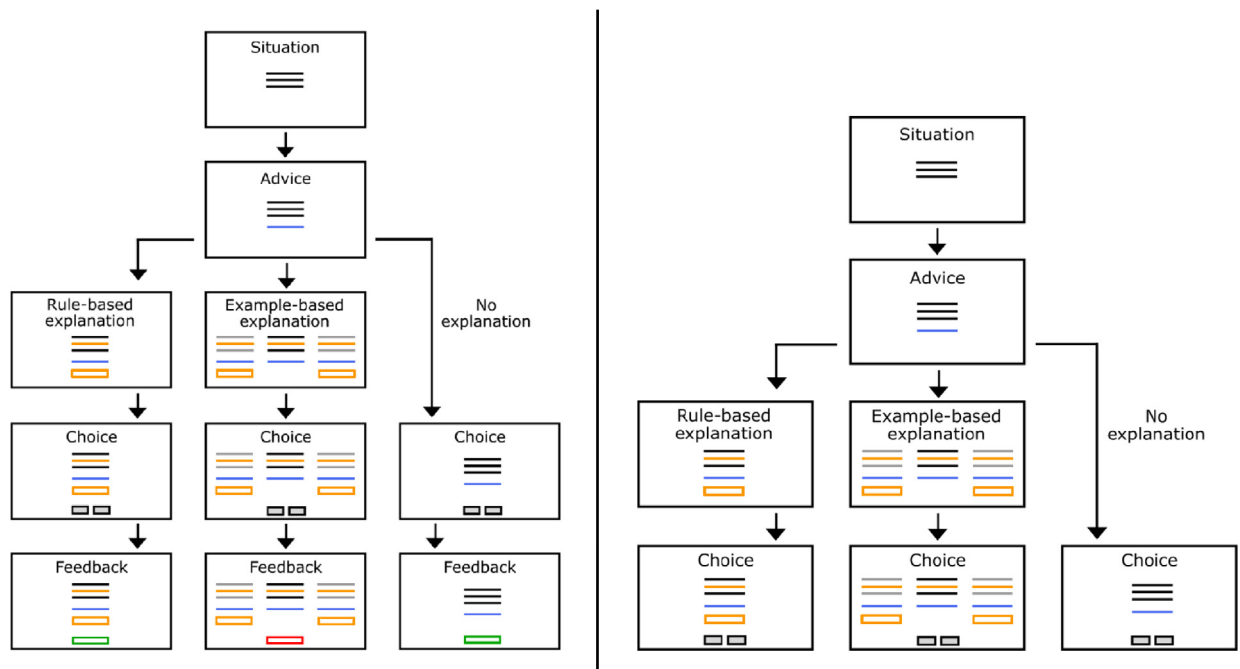
Fig. 5. A schematic overview of the learning (left) and testing (right) block in Experiment II.

measurement of *advice prediction*. The measurement *decisive factor identification* was formed by the subsequent question to select a single factor from a situation description that they believed was decisive for the predicted system advice.

A third, self-reported measurement was conducted in the post-questionnaire, which contained an eight-item question-naire based on a 7-point Likert scale. These items formed the measurement of *perceived system understanding*. The questions were asked without mentioning the term explanation and simply addressed 'system output'. The amount of eight items was deemed necessary, to obtain a measurement less dependent on the formulation of one item.

6.2. Experiment II: Persuasive power and task performance

The purpose of *Experiment II* was to measure the effects of *rule-based* and *example-based* explanations on persuasive power and task performance, and to compare these to each other and to the control group with *no explanation*. Fig. 5 provides an overview of the learning and testing blocks of this experiment. The learning block was similar to that of the first experiment: a situation was shown, containing three factors from Table 1. In the experimental groups, the situation was followed by an advice and explanation. Next, the participant was asked to make a decision on the insulin dose. After this point, the learning block differed from the learning block in the first experiment: the participant's decision was followed with feedback on its correctness. In 12 of the 18 randomly ordered trials of this learning block (66%), the system's advice was correct. In the six other trials, the advice was incorrect. Through this feedback, participants learned that the system's advice could be incorrect and in which situations. Instead of following the ground truth rule set (from *Experiment I*), this system followed a second, partially correct set of rules, as shown in Table 1.

The testing block contained 30 trials, also presented in random order, in which a presented situation was followed by the system's advice and explanation. Next, participants had to choose which insulin dose was correct based on the system's advice, explanation and gained knowledge of when the system is incorrect. Persuasive power was operationalized as the number of times a participant followed the advice, independent of whether it was correct or not. Task performance was represented by the number of times a correct decision was made. The former reflected how persuasive the advice and explanation was, even when participants experienced system errors. The latter reflected how well participants were able to understand when the system makes errors and compensate accordingly in their decision.

Also in this experiment, a self-reported measurement with eight 7-point Likert scale questions was performed. It measured the participant's subjective sense of their ability to estimate when the system was correct.

6.3. Participants

In *Experiment I*, 45 participants took part, of which 21 female and 24 male, aged between 18 and 64 years old ($M = 44.2 \pm 16.8$). Their education levels varied from lower vocational to university education. In *Experiment II* 45 dif-ferent participants took part, of which 31 female and 14 male, aged between 18 and 61 years old ($M = 36.5 \pm 14.5$). Their

education levels varied from secondary vocational to university education. Participants were recruited from a participant database at TNO Soesterberg (NL) as well as via advertisements in Utrecht University (NL) buildings and on social media. Participants received a compensation of 20,- euro and their travel costs were reimbursed. Both samples aimed to represent the entire Dutch population and as such the entire range of potential DMT1 patients, hence the wide age and educational ranges.

The inclusion criteria were as follows: not diabetic, no close relatives or friends with diabetes, and no extensive knowledge of diabetes through work or education. General criteria were Dutch native speaking, good or corrected eyesight, and basic experience using computers. These inclusion criteria were verified in the pre-questionnaire. A total of 16 participants reported a close relative or friend with diabetes and one participant had experience with diabetes through work, despite clear inclusion instructions beforehand. After careful inspection of their answers, none were excluded because their answers on diabetes questions in the pre-questionnaire were not more accurate or elaborate than others. From this we concluded that their knowledge of diabetes was unlikely to influence the results.

7. Data analysis

Statistical tests were conducted using SPSS Statistics 22. An alpha level of 0.05 was used for all statistical tests.

The data from the behavioral measures in *Experiment I* were analyzed using a one-way Multivariate Analysis of Variance (MANOVA) with explanation style (*rule-based*, *example-based* or *no explanation*) as the independent between-subjects variable and *advice prediction* and *decisive factor identification* as dependent variables. The reason for a one-way MANOVA was the multivariate operationalization of a single construct, understanding [66]. Cronbach's Alpha was used to assess the internal consistency of the self-reported measurement for *perceived system understanding* from the post-questionnaire. Subsequently, a one-way Analysis of Variance (ANOVA) was conducted with the mean rating on this questionnaire as dependent variable and the explanation style as independent variable. Finally, the relation between the two behavioral and the self-reported measurements was examined with Pearson's product-moment correlations.

For *Experiment II* two one-way ANOVA's were performed. The first ANOVA had the explanation style (*rule-based*, *example-based* or *no explanation*) as independent variable and the *number of times the advice was copied* as dependent variable. The second ANOVA also had explanation style as independent variable, but the *number of correct decisions* as dependent variable. The internal consistency of the self-reported measurement of *perceived prediction of advice correctness* from the post-questionnaire was assessed with Cronbach's Alpha and analyzed with a one-way ANOVA. Explanation style was the independent and the mean rating on the questionnaire the dependent variable. The presence of correlations between the behavioral and the self-reported measurements was assessed with Pearson's product-moment correlations. Detected outliers were excluded from the analysis.

8. Results

8.1. Experiment I: System understanding

The purpose of *Experiment I* was to measure gained system understanding when a system provides a rule- or example-based explanation, compared to no explanation. This was measured with two behavioral measures and one self-reported measure.

Fig. 6 shows the results on the two behavioral measures: correct advice prediction in novel situations and correct identification of the system's decisive factor. A one-way MANOVA with Wilks' lambda indicated a significant main effect of explanation style on both measurements ($F(4, 82) = 6.675$, $p < 0.001$, $\Delta = .450$, $\eta_p^2 = .246$). Further analysis revealed a significant effect for explanation style on factor identification ($F(2, 42) = 14.816$, $p < 0.001$, $\eta_p^2 = .414$), but not for advice prediction ($F(2, 42) = 14.816$, $p = .264$, $\eta_p^2 = .414$). One assumption of a one-way MANOVA was violated, as the linear relationships between the two dependent variables and each explanation style was weak. This was indicated by Pearson's product-moment correlations for the rule-based ($r = .487$, $p = .066$), example-based ($r = -.179$, $p = .522$) and no explanation ($r = .134$, $p = .636$) groups. Some caution is needed in interpreting these results, as this lack of significant correlations shows a potential lack of statistical power. Further post-hoc analysis showed a significant difference in factor identification in favor of rule-based explanations compared to example-based explanations and no explanations ($p < 0.001$). No significant difference between example-based explanations and no explanation was found ($p = .796$).

Fig. 7 shows the results on the self-reported measure of system understanding. The consistency between the different items in the measure was very high, as reflected by Cronbach's alpha ($\alpha = .904$). The mean rating over all eight items was used as the participant's subjective rating of system understanding. A one-way ANOVA showed a significant main effect of explanation style on this rating ($F(2, 41) = 7.222$, $p = .002$, $p\eta_p^2 = .261$). Two assumptions of a one-way ANOVA were violated. First, the rule-based explanations group had one outlier, of which the exclusion did not affect the analysis in any way. The results after removal of this outlier are reported. Second, Levene's test was not significant ($p = .017$) signaling inequality between group variances. However, ANOVA is robust against the variance homogeneity violation with equal group sizes [67,68]. Further post-hoc tests revealed that only rule-based explanations caused a significantly higher self-reported understanding compared to no explanations ($p = .001$). No significant difference was found for example-based explanations with no explanations ($p = .283$) and with rule-based explanations ($p = .072$).
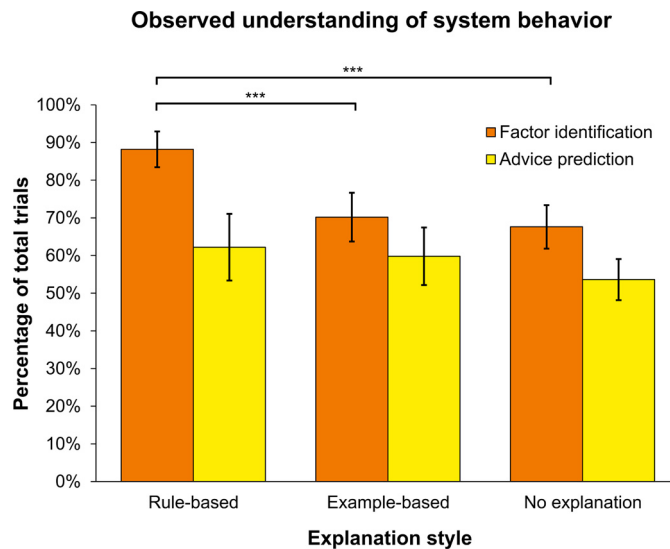
Observed understanding of system behavior

Fig. 6. Bar plot of the mean percentages of correct prediction of the system's advice and correct identification of the decisive factor for that advice. Values are relative to the total of 30 randomized trials in Experiment I. The error bars represent a 95% confidence interval. *Note;* ***$p < 0.001$.

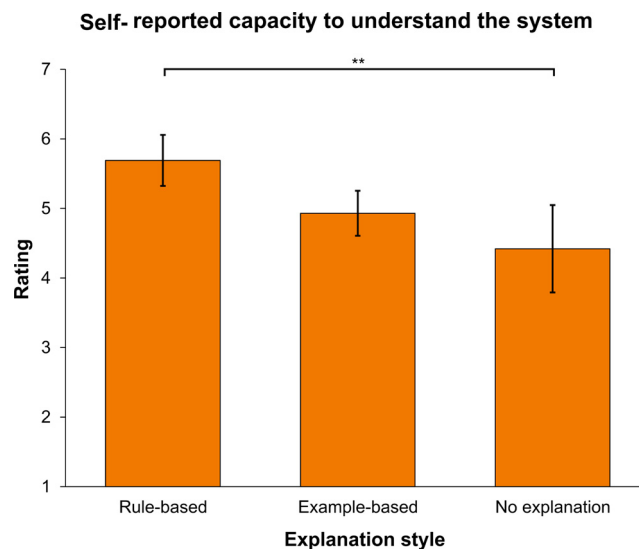**Self- reported capacity to understand the system**

Fig. 7. Bar plot of the mean self-reported system understanding. All values are on a 7-point Likert scale and error bars represent a 95% confidence interval. *Note;* ** $p < 0.01$.

Finally, Fig. 8 shows a scatter plot between both behavioral measures and the self-reported measure. Pearson's product-moment analysis revealed no significant correlations between self-reported understanding and advice prediction ($r = -.007$, $p = .965$), not within the rule-based explanation group ($r = -.462$, $p = .129$), the example-based explanation group ($r = -.098$, $p = .729$), nor the no explanation group ($r = .001$, $p = .996$). Similar results were found for the correlation between self-reported understanding and factor identification ($r = .192$, $p = .211$) and for the separate groups of rule-based explanations ($r = -.124$, $p = .673$), example-based explanations ($r = .057$, $p = .840$) and no explanations ($r = -.394$, $p = .146$).

8.2. Experiment II: Persuasive power and task performance

The purpose of *Experiment II* was to measure a participant's ability to use a decision support system appropriately when it provides a rule- or example-based explanation, compared with no explanation. This was measured with one behavioral and one self-reported measurement. In addition, we measured the persuasiveness of the system for each explanation style, compared to no explanations. This was assessed with one behavioral measure.
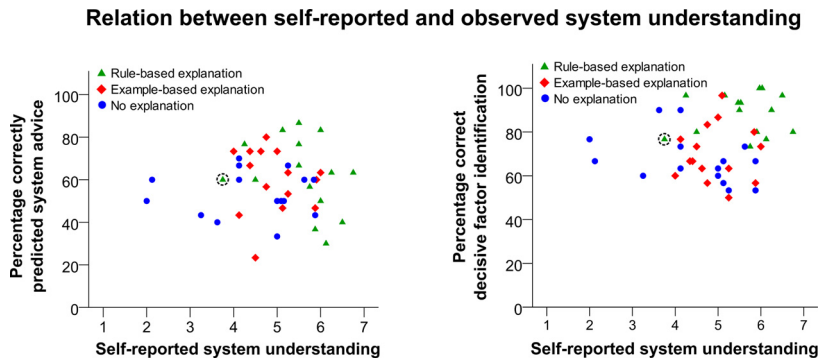
Relation between self-reported and observed system understanding

Fig. 8. Scatter plots displaying the relation between advice prediction (left) and decisive factor identification (right) with self-reported understanding. Outliers are circled.
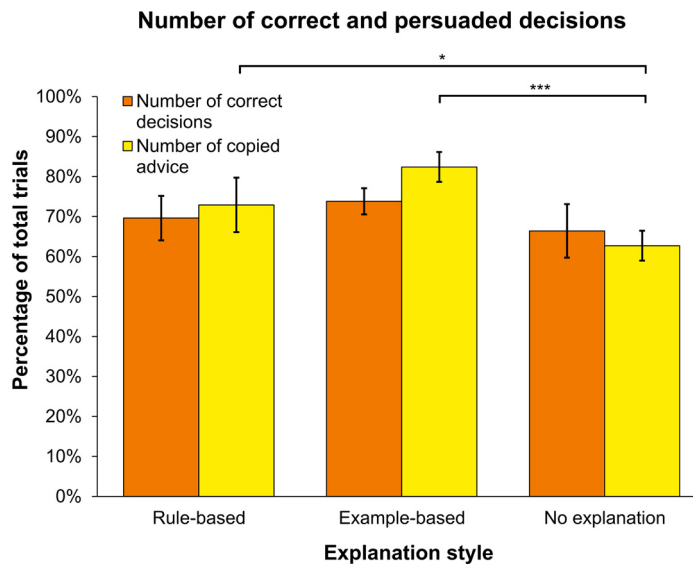
Number of correct and persuaded decisions

Fig. 9. Bar plot displaying task performance (the mean percentage of correct decisions) and persuasive power (the mean percentage of decisions following the system's advice independent of correctness). Error bars represent a 95% confidence interval. *Note;* $^*p < 0.05$, $^{***}p < 0.001$.

Fig. 9 shows the results of the behavioral measure for task performance, as reflected by the user's decision accuracy. A one-way ANOVA showed no significant differences ($F(2, 41) = 1.716$, $p = .192$, $\eta_p^2 = .077$). Two violations of ANOVA were discovered. There was one outlier in the example-based explanations, with 93.3% accuracy (1 error). Removal of the outlier did not affect the analysis. Levene's test showed there was no homogeneity of variances ($p = .007$), however ANOVA is believed to be robust against this under equal group sizes [67,68].

Fig. 9 shows the results of the behavioral measure for persuasiveness, i.e. the number times system advice was followed. Note that in *Experiment II* the system's accuracy was 66.7%. Thus, following the advice in a higher percentage of cases denotes an adverse amount of persuasion. A one-way ANOVA showed that explanation style had a significant effect on following the system's advice ($F(2, 41) = 11.593$, $p < .001$, $\eta_p^2 = .361$). Further analysis revealed that participants with no explanation followed the system's advice significantly less than those with rule-based ($p = .049$) and example-based explanations ($p < .001$). However, there was no significant difference between the two explanation styles ($p = .068$). One outlier violated the assumptions of an ANOVA. One participant in the rule-based explanation group followed the system's advice only 33.3% of the time. Its exclusion affected the outcomes of the ANOVA and the results after exclusion are reported.

Fig. 10 displays the self-reported capacity to predict correctness, operationalized by a rating how well participants thought they were able to predict when system advice was correct or not. The consistency of the eight 7-point Likert scale questions was high according to Cronbach's Alpha ($\alpha = .820$). Therefore, we took the mean rating of all questions as an estimate of participants' performance estimation. A one-way ANOVA was performed, revealing no significant differences ($F(2, 41) = 2.848$, $p = .069$, $\eta_p^2 = .122$). One outlier from the rule-based explanation group was found, its removal did not affect the analysis.

A correlation analysis was performed between the self-reported prediction of advice correctness and the behavioral measurement of making the correct decision, two measurements of task performance. The accompanying scatter plot is shown in Fig. 11. A Pearson's product-moment correlation revealed no significant correlation between the self-reported and
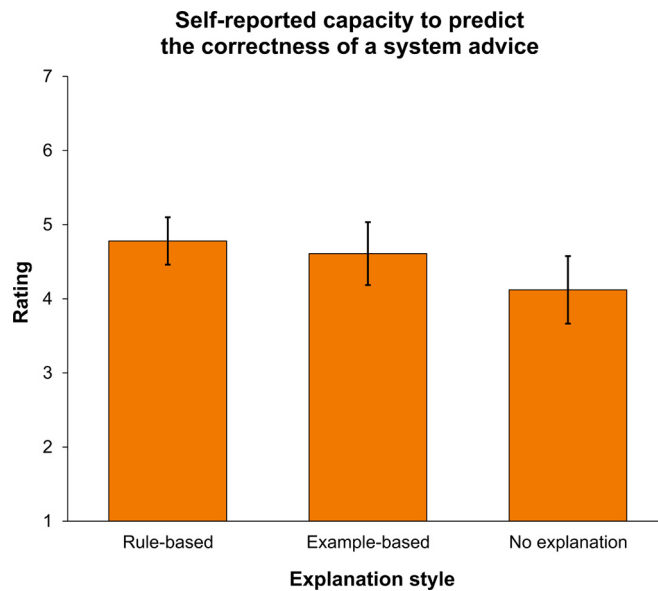
**Self-reported capacity to predict
the correctness of a system advice**

Fig. 10. Bar plot of the mean self-reported system performance estimation. All values are on a 7-point Likert scale and error bars represent a 95% confidence interval.
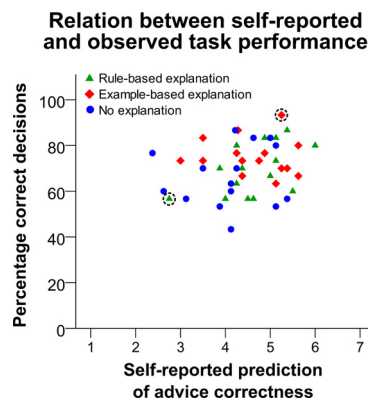
**Relation between self-reported
and observed task performance**

Fig. 11. Scatter plot displaying the relation between the number of correct decisions made and the self-reported capacity to predict advice correctness. Outliers are circled.

behavioral measure ($r = .146$, $p = .350$). Also, there were no significant correlations in the rule-based ($r = .411$, $p = .144$) and example-based explanation ($r = -.347$, $p = .225$) groups, nor in the no explanation group ($r = .102$, $p = .718$). Both outliers from each measurement were removed in this analysis and did not affect the significance.

8.3. Usability and biases

A usability questionnaire was used to evaluate whether there were differences in usability between the two explanation styles, as this could influence the results. The questionnaire contained five questions on a 100-point scale about readability, organization of information, language, images and color. The consistency between the five questions was relatively high, as revealed by a Cronbach's Alpha test ($\alpha = .722$). Fig. 12 shows the mean ratings for each question, broken down by explanation style (*rule-based*, *example-based*, *no explanation*). No statistical analysis was performed, as this questionnaire only functioned as a check for potential usability confounds in the experiment.

In addition to the ratings, participants were asked about the positive and negative usability aspects of the system in two open questions. Common positive descriptions included "clear", "well-arranged", "clear and simple icons" and "understandable language". Although not many participants had negative remarks, most addressed insufficient visual contrast due to the colors used. Unique to the example-based explanations participant group were remarks about a lack of concise and well-arranged information.

In the control questionnaire we asked participants to give an estimate of the overall system's accuracy. This was to validate any potential overly positive or negative trust bias towards the system. In *Experiment I* the system was 100% accurate, but this was unknown to the participants since there was no feedback on correctness included. Nonetheless, estimates
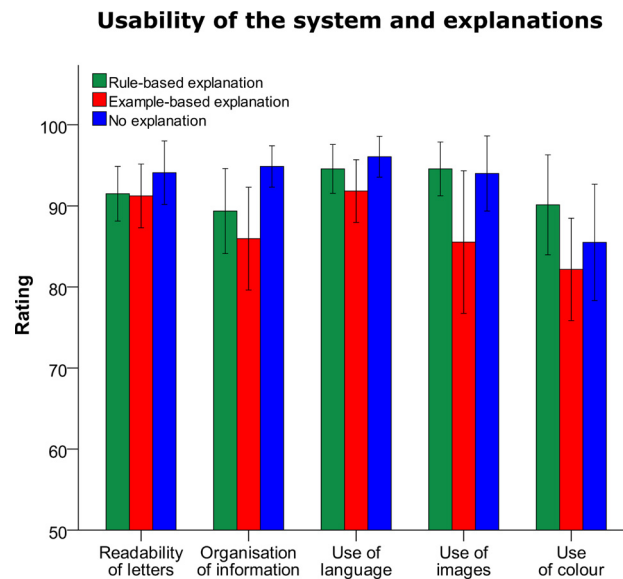
Usability of the system and explanations

Fig. 12. The mean ratings on the usability questions, displayed by explanation style. The error bars represent a 95% confidence interval.

ranged from 30% to 90% ($\mu = 75.2\%$, $\sigma = 12.8\%$). This meant that all participants believed the system to make errors based on no information. In *Experiment II* the system's accuracy was 66.7%. Participants experienced this due to the feedback on made decisions in the learning block. Estimates ranged between 50% and 95% ($\mu = 74.8\%$, $\sigma = 8.8\%$), indicating that on average, system accuracy was overestimated.

After the experiment, brief discussions with participants revealed additional perspectives. Several participants from the no explanation group wished the system could give an explanation for its advice. One participant expressed a need for knowing the system's rules governing the system's advice. In the two explanation groups, participants experienced the explanations as useful. Rules were valued for there explicitness, whereas examples were viewed as inciting trust. However, in the two explanation groups several participants found it unclear what the highlight of a factor (see Fig. 3) meant. Several participants also mentioned that, although useful, the explanations lacked a causal rationale.

9. Discussion

Below we discuss the results from both experiments in detail and relate them to our theory presented in Section 5.

9.1. Experiment I: System understanding

Experiment I measured the participant's capacity to understand how and when the system provided a specific advice. This construct was operationalized in three measurements: *decisive factor identification*, *advice prediction* and *perceived system understanding*. We hypothesized that participants receiving *contrastive rule-based explanations* would score best on all three measurements. *Contrastive example-based explanations* were only expected to improve understanding slightly more than *no-explanations* (see Fig. 2).

The results from our evaluation support these hypotheses in part. First, *rule-based explanations* indeed seem to allow participants to more accurately identify the factor from a situation that was decisive in the system's advice. However, *rule-based* nor *example-based explanation* allowed participants to learn to predict system behavior. The *rule-based explanations* however, did cause to participants to think that they better understood the system compared to *example-based* and *no explanations*. The *example-based explanations* only showed a small and insignificant increase in perceived system understanding. It is important to note that there was no correlation between the self-reported measurement of understanding and the behavioral measurements of understanding. This shows that participants had a perception of understanding that differed from the understanding as measured with *factor identification* and *advice prediction*.

Close inspection of the results showed two potential causes for the lack of support for our hypotheses. The first reason might be because the described DMT1 situations and accompanying system advice was too intuitive. This is supported by the fact that participants with *no explanation* were already quite adapt in *identifying decisive factors* (nearly 70% compared to 33% chance). The second reason we inferred from open discussions with participants after the experiment. Most participants who received either explanation style mentioned difficulty in applying and generalizing the knowledge from the explanations to novel situations. Several participants even expressed the desire to know the rationale of why a certain rule or behavior occurred. This is in line with the theory that explanations should convey specific causal relations obtained from an overall causal model describing the behavior of the system, instead of just factual correlations between system input and output.

If we generalize these results to the field of XAI, we have shown that *contrastive rule-based explanations* as "if... then..." statements are not sufficient to predict system behavior. However, such explanations are capable of educating a user to identify which factors would play a decisive role in system advice given a specific situation. Also, such explanations seem to provide the user with the perception that (s)he is better capable of understanding the system. The *contrastive example-based explanations* however showed no improvement on observed or self-reported understanding. This experiment illustrated the need for explanations that provide more causal information, instead of solely information depicting system input and output correlations. Furthermore, we illustrated that self-reported and behavioral measurements of understanding may not correlate, underlining the need for (a combination of) measures that accurately and reliably measure the intended construct.

9.2. Experiment II: Persuasive power and task performance

In *Experiment II* we investigated the extent to which an explanation increases the persuasiveness of an advice, as well as the explanation's effect on task performance. The persuasive power of an explanation was operationalized with the *number of times the advice was copied*. Task performance was represented by the *number of correct decisions* and the self-reported *prediction of advice correctness*. We hypothesized that especially *contrastive example-based explanations* would increase persuasive power, while these in turn would lower actual task performance. In contrast, the understanding participants gained from *rule-based explanations* was expected to cause an increase in task performance (see Fig. 2).

Both *contrastive rule-based* and *example-based* explanations showed more persuasive power than when *no explanation* was given. The *example-based explanations* also showed slightly more persuasive power than the *rule-based explanations*, but this difference was not significant. These results partly support our theory about persuasive power, as they illustrate that explanations persuade users to follow a system's advice more often. These results however, do not support that *example-based explanations* are much more persuasive than *rule-based explanations*.

With respect to task performance, we saw that explanations caused small but insignificant improvements on both behavioral and self-reported data. In fact, the *example-based explanations* showed the highest (but still insignificant) improvement. Due to a lack of statistical evidence not much can be inferred from this, and further evaluation is required.

Similar to *Experiment I* we found a lack of correlation between reports of participants' perception of *predicting advice correctness*, and the *number of correct decisions*. In other words, these measures do not seem to measure the same construct. An explanation could be that participants were unable to estimate their own capacity of predicting the correctness of advice.

We have shown that providing an explanation with an advice results in users following that advice more often, even when incorrect. In addition, there was a suggestion that explanations also improve task performance, especially *contrastive example-based explanations*. However, these effects were marginal and not significant. These results underline the need in the field of XAI to take a different stance on which explanations should be generated. Two common styles of explanations answering a contrasting question did not appear to increase task performance, an effect often attributed to such explanations within the field.

10. Limitations

This study has several limitations that warrant caution in generalizing the results to other use cases or to the field of XAI in general. The first set of limitations is related to the selected use case of aided DMT1 self-management. This use case falls into the category 'simplified' from Doshi-Velez and Kim [4] as it approximates a realistic use case. However, two major aspects differ from the real-life situation. First, we recruited healthy participants who had to empathize with a DMT1 patient, instead of actual DMT1 patients. Nevertheless, participants were sampled from the entire Dutch population, resulting in a wide variety of ages and education levels. These choices allowed us to measure the effects of the explanation types without focusing on a specific demographic or having to compensate for varying domain knowledge in DMT1 participants. Second, the system itself was fictitious and followed a pre-determined set of rules rather than comprising the full complexity of a realistic system. These two simplifications prevent us to generalize the results and to apply our conclusions to construct an actual system for aiding DMT1 patients in self-management. However, this was not the purpose of this study. Instead, we aimed to evaluate whether the supposed effects of two often cited explanations styles were warranted. We believe the selected use case allowed us to do so, as it gave both context as well as motivation for the users to understand explanations. Also, laymen were chosen opposed to DMT1 patients to mitigate any difference in diabetes knowledge and misconceptions, which can vary greatly between patients (e.g. see [55]). Of course, future research specifically targeted at the development of a DSS for DMT1 self-management should include DMT1 patients as participants.

The second set of limitations is related to suspected confounds in the experiment. A brief usability questionnaire showed that participants held an overall positive bias towards the system, whether an explanation was provided or not. In addition this questionnaire showed that participants' perception of the organization of the information was not always positive. Hence, a potential limitation lies in the way the explanations were presented. Also, surprisingly, in *Experiment I* participants attributed a low performance to the system, while they had no information to do so. In *Experiment II* however, participants tended to slightly overestimate the system's actual performance. This occurred independent of the explanation style. This shows that the participants could have had a natural tendency to distrust the system's advice. This may have affected the self-reported results.

Finally, a few limitations arose from the design of both experiments. The results for the example-based explanations could have been different with a longer learning block, as it takes time to infer decision boundaries from examples. Also, both testing blocks were relatively long, which could have caused participants to continue learning about the system while we were measuring their understanding. We did not perform any analyses on this, as it would add another level of complexity to the design. Hence, we cannot say for certain that the learning block was of sufficient length to allow participants to learn enough from the explanations. However, if this was the case, we believe that prolonging the learning block would have resulted in even stronger effects. Lastly, due to the choice of different participant groups for both experiments, we could only draw limited conclusions on the relation between the understanding on the one hand and task performance and persuasiveness on the other hand. However, we selected this approach instead of combining the constructs in a single experiment with a within-subject design, to avoid learning effects not sufficiently compensated through randomizing the understanding and task performance/persuasion blocks.

11. Conclusion

A lack of user evaluations characterizes the field of Explainable Artificial Intelligence (XAI). A contribution of this paper was to provide a set of recommendations for future user evaluations. Practical recommendations were given for XAI researchers unfamiliar with user evaluations. These addressed the evaluation's constructs and their relations, the selection of a use case and the experimental context, and suitable measurements to operationalize the constructs in the evaluation. These recommendations originated from our experience designing an extensive user evaluation. Our second contribution was to evaluate the effects of *contrastive rule-based* and *contrastive example-based explanations* on the participant's understanding of system behavior, persuasive power of the system's advice when combined with an explanation, and task performance. The evaluation took place in a decision-support context where users were aided in choosing the appropriate dose of insulin to mitigate the effects of diabetes mellitus type 1.

Results showed that *contrastive rule-based explanations* allowed participants to correctly identify the situational factor that played a decisive role in a system's advice. Neither *example-based* or *rule-based explanations* enabled participants to correctly predict the system's advice in novel situations, nor did they improve task performance. However, both explanation styles did cause participants to follow the system's advice more often, even when this advice was incorrect. This shows that both *rules* and *examples* that answer a contrastive question are not sufficient on their own to improve users' understanding or task performance. We believe that the main reason for this is that these explanations lack a clarification of the underlying rationale of system behavior.

Future work will focus on the evaluation of a combined explanation style provided in interactive form, to assess whether this interactive form helps users to learn a system's underlying rationale. As an extension, potential methods will be researched that can generate causal reasoning traces, rather than decision boundaries, to expose the behavior rationale directly. In addition, future research may focus on similar studies with actual diabetes patients to study explanation effects in potentially homogeneous groups (e.g. effects of age, domain knowledge, etc.). Finally, during the design and analysis of this user evaluation we discovered a need for validated and reliable measurements. We will continue to use different types of measurements to measure constructs in a valid and reliable way in future user evaluations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

References

[1] M.M. De Graaf, B.F. Malle, How people explain action (and autonomous intelligent systems should too), in: 2017 AAAI Fall Symposium Series, 2017.
[2] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267C (2019) 1–38.
[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2019) 93.
[4] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608.
[5] S. Anjomshoae, A. Najjar, D. Calvaresi, K. Främling, Explainable agents and robots: results from a systematic literature review, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1078–1088.
[6] T. Miller, Contrastive explanation: a structural-model approach, arXiv preprint arXiv:1811.03163.
[7] R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: challenges and prospects, arXiv preprint arXiv:1812.04608.
[8] E.J. de Visser, M.M. Peeters, M.F. Jung, S. Kohn, T.H. Shaw, R. Pak, M.A. Neerincx, Towards a theory of longitudinal trust calibration in human–robot teams, Int. J. Soc. Robot. 12 (2) (2020) 459–478.

[9] I. Contreras, J. Vehi, Artificial intelligence for diabetes management and decision support: literature review, J. Med. Internet Res. 20 (5) (2018) e10775.
[10] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, I. Chouvarda, Machine learning and data mining methods in diabetes research, Comput. Struct. Biotechnol. J. 15 (2017) 104–116.
[11] M. Neerincx, W. Vught, O. Blanson Henkemans, E. Oleari, J. Broekens, R. Peters, F. Kaptein, Y. Demiris, B. Kiefer, D. Fumagalli, et al., Socio-cognitive engineering of a robotic partner for child's diabetes self-management, Front. Robot. AI 6 (2019) 118.
[12] B. Hayes, J.A. Shah, Improving robot controller transparency through autonomous policy explanation, in: 2017 12th ACM/IEEE International Conference on Human-Robot Interaction, HRI, IEEE, 2017, pp. 303–312.
[13] T. Chakraborti, A. Kulkarni, S. Sreedharan, D.E. Smith, S. Kambhampati, Explicability? Legibility? Predictability? Transparency? Privacy? Security? The emerging landscape of interpretable agent behavior, in: Proceedings of the International Conference on Automated Planning and Scheduling, vol. 29, 2019, pp. 86–96.
[14] J.E. Mercado, M.A. Rupp, J.Y. Chen, M.J. Barnes, D. Barber, K. Procci, Intelligent agent transparency in human–agent teaming for multi-uxv management, Hum. Factors 58 (3) (2016) 401–415.
[15] K. Stubbs, P.J. Hinds, D. Wettergreen, Autonomy and common ground in human-robot interaction: a field study, IEEE Intell. Syst. 22 (2) (2007) 42–50.
[16] T. Kulesza, M. Burnett, W.-K. Wong, S. Stumpf, Principles of explanatory debugging to personalize interactive machine learning, in: Proceedings of the 20th International Conference on Intelligent User Interfaces, ACM, 2015, pp. 126–137.
[17] O. Biran, C. Cotton, Explanation and justification in machine learning: a survey, in: IJCAI-17 Workshop on Explainable AI (XAI), vol. 8, 2017, p. 1.
[18] Z.C. Lipton, The mythos of model interpretability, arXiv preprint arXiv:1606.03490.
[19] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
[20] M. Bilgic, R.J. Mooney, Explaining recommendations: satisfaction vs. promotion, in: Beyond Personalization Workshop, IUI, vol. 5, 2005, p. 153.
[21] U. Ehsan, B. Harrison, L. Chan, M.O. Riedl, Rationalization: a neural machine translation approach to generating natural language explanations, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, ACM, 2018, pp. 81–87.
[22] L.A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darrell, Generating visual explanations, in: European Conference on Computer Vision, Springer, 2016, pp. 3–19.
[23] J.L. Herlocker, J.A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, ACM, 2000, pp. 241–250.
[24] L.R. Ye, P.E. Johnson, The impact of explanation facilities on user acceptance of expert systems advice, MIS Q. (1995) 157–172.
[25] J. Zhou, Z. Li, H. Hu, K. Yu, F. Chen, Z. Li, Y. Wang, Effects of influence on user trust in predictive decision making, in: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–6.
[26] S. Berkovsky, R. Taib, D. Conway, How to recommend?: User trust factors in movie recommender systems, in: Proceedings of the 22nd International Conference on Intelligent User Interfaces, ACM, 2017, pp. 287–300.
[27] D. Holliday, S. Wilson, S. Stumpf, User trust in intelligent systems: a journey over time, in: Proceedings of the 21st International Conference on Intelligent User Interfaces, ACM, 2016, pp. 164–168.
[28] F. Nothdurft, T. Heinroth, W. Minker, The impact of explanation dialogues on human-computer trust, in: International Conference on Human-Computer Interaction, Springer, 2013, pp. 59–67.
[29] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, F. Doshi-Velez, An evaluation of the human-interpretability of explanation, arXiv preprint arXiv:1902.00006.
[30] M. Joppe, 2000. The Research Process. Retrieved February 25, 1998.
[31] E.A. Drost, et al., Validity and reliability in social science research, Educ. Res. Perspect. 38 (1) (2011) 105.
[32] J. Kirk, M.L. Miller, M.L. Miller, Reliability and Validity in Qualitative Research, vol. 1, Sage, 1986.
[33] P. Lipton, Contrastive explanation, R. Inst. Philos. Suppl. 27 (1990) 247–266.
[34] B.Y. Lim, A.K. Dey, Assessing demand for intelligibility in context-aware applications, in: Proceedings of the 11th International Conference on Ubiquitous Computing, ACM, 2009, pp. 195–204.
[35] L.K. Branting, Building explanations from rules and structured cases, Int. J. Man-Mach. Stud. 34 (6) (1991) 797–837.
[36] J. van der Waa, M. Robeer, J. van Diggelen, M. Brinkhuis, M. Neerincx, Contrastive explanations with local foil trees, arXiv preprint arXiv:1806.07470.
[37] F. Wang, C. Rudin, Falling rule lists, in: Artificial Intelligence and Statistics, 2015, pp. 1013–1022.
[38] N. Barakat, J. Diederich, Eclectic rule-extraction from support vector machines, Int. J. Comput. Intell. 2 (1) (2005) 59–62.
[39] A. Newell, H.A. Simon, et al., Human Problem Solving, vol. 104, Prentice-Hall, Englewood Cliffs, NJ, 1972.
[40] M.T. Chi, M. Bassok, M.W. Lewis, P. Reimann, R. Glaser, Self-explanations: how students study and use examples in learning to solve problems, Cogn. Sci. 13 (2) (1989) 145–182.
[41] A. Renkl, Worked-out examples: instructional explanations support learning by self-explanations, Learn. Instr. 12 (5) (2002) 529–556.
[42] I. Peled, O. Zaslavsky, Counter-examples that (only) prove and counter-examples that (also) explain, Focus Learn. Probl. Math. 19 (3) (1997) 49–61.
[43] A. Adhikari, D.M. Tax, R. Satta, M. Faeth, Leafage: example-based and feature importance-based explanations for black-box ml models, in: 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2019, pp. 1–7.
[44] J. Bien, R. Tibshirani, et al., Prototype selection for interpretable classification, Ann. Appl. Stat. 5 (4) (2011) 2403–2424.
[45] B. Kim, C. Rudin, J.A. Shah, The bayesian case model: a generative approach for case-based reasoning and prototype classification, in: Advances in Neural Information Processing Systems, 2014, pp. 1952–1960.
[46] B. Kim, R. Khanna, O.O. Koyejo, Examples are not enough, learn to criticize! Criticism for interpretability, in: Advances in Neural Information Processing Systems, 2016, pp. 2280–2288.
[47] R.K. Atkinson, Optimizing learning from examples using animated pedagogical agents, J. Educ. Psychol. 94 (2) (2002) 416.
[48] M.J. Pazzani, Representation of electronic mail filtering profiles: a user study, in: Proceedings of the 5th International Conference on Intelligent User Interfaces, 2000, pp. 202–206.
[49] S. Stumpf, V. Rajaram, L. Li, W.-K. Wong, M. Burnett, T. Dietterich, E. Sullivan, J. Herlocker, Interacting meaningfully with machine learning systems: three experiments, Int. J. Hum.-Comput. Stud. 67 (8) (2009) 639–662.
[50] A. Bussone, S. Stumpf, D. O'Sullivan, The role of explanations on trust and reliance in clinical decision support systems, in: 2015 International Conference on Healthcare Informatics, IEEE, 2015, pp. 160–169.
[51] B.Y. Lim, A.K. Dey, D. Avrahami, Why and why not explanations improve the intelligibility of context-aware intelligent systems, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2009, pp. 2119–2128.
[52] J. Pearl, et al., Causal inference in statistics: an overview, Stat. Surv. 3 (2009) 96–146.
[53] V.O. Mittal, C.L. Paris, Generating explanations in context: the system perspective, Expert Syst. Appl. 8 (4) (1995) 491–503.
[54] N.J. Cooke, S.M. Shope, L. Schiflett, E. Salas, M. Coovert, Designing a synthetic task environment, in: Scaled Worlds: Development, Validation, and Application, 2004, pp. 263–278.
[55] V.U. Odili, P.D. Isiboge, A. Eregie, Patients' knowledge of diabetes mellitus in a Nigerian city, Trop. J. Pharm. Res. 10 (5) (2011) 637–642.
[56] G. Paolacci, J. Chandler, P.G. Ipeirotis, Running experiments on Amazon mechanical turk, Judgm. Decis. Mak. 5 (5) (2010) 411–419.
[57] A. Papenmeier, G. Englebienne, C. Seifert, How model accuracy and explanation fidelity influence user trust, arXiv preprint arXiv:1907.12652.

[58] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, How can we fool LIME and SHAP? Adversarial attacks on post hoc explanation methods, arXiv preprint arXiv:1911.02508.

[59] Y. Zhang, K. Song, Y. Sun, S. Tan, M. Udell, "Why should you trust my explanation?" understanding uncertainty in lime explanations, arXiv preprint arXiv:1904.12991.

[60] V. Buch, G. Varughese, M. Maruthappu, Artificial intelligence in diabetes care, Diabet. Med. 35 (4) (2018) 495–497.

[61] M. Reddy, S. Rilstone, P. Cooper, N.S. Oliver, Type 1 diabetes in adults: supporting self management, BMJ, Br. Med. J. 352 (2016) I998.

[62] I. Bosch, Het Groot Diabetesboek, Mension B.V., 2013.

[63] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (4) (2019) e1312.

[64] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215.

[65] R.C. Bogdan, S.K. Biklen, Qualitative Research in (Validation) and Qualitative (Inquiry) Studies, Allyn & Bacon, 2006.

[66] C.J. Huberty, J.D. Morris, Multivariate analysis versus multiple univariate analyses, Psychol. Bull. 105 (1989) 302–308.

[67] G. Keppel, Design and Analysis: A Researcher's Handbook, Prentice-Hall, Inc., 1991.

[68] E. Yigit, F. Gokpinar, A simulation study on tests for one-way ANOVA under the unequal variance assumption, Commun. Fac. Sci. Univ. Ank. Sér. A 1 (2010) 15–34.